Data Center / Cloud

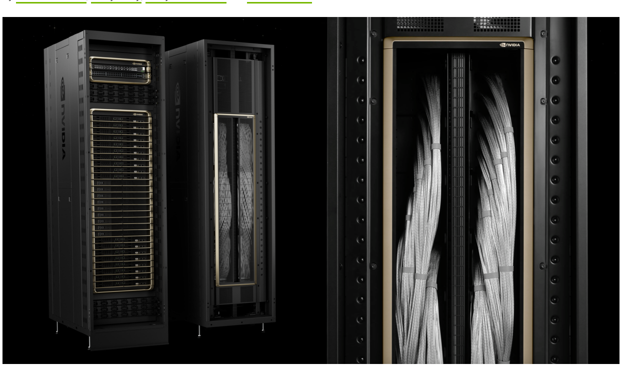
English

NVIDIA GB200 NVL72 Delivers Trillion-Parameter LLM Training and Real-Time Inference

Mar 18, 2024

+28 Like Discuss (5)

By Ivan Goldwasser, Harry Petty, Pradyumna Desale and Kirthi Devleker



What is the interest in trillion-parameter models? We know many of the use cases today and interest is growing due to the promise of an increased capacity for:

- Natural language processing tasks like translation, question answering, abstraction, and fluency.
- Holding longer-term context and conversational ability.
- Multimodal applications combining language, vision, and speech.
- Creative applications like storytelling, poetry generation, and code generation.
- Scientific applications, such as protein folding predictions and drug discovery.
- Personalization, with the ability to develop a consistent personality and remember user context.

The benefits are great, but training and deploying large models can be computationally expensive and resource-intensive. Computationally efficient, cost-effective, and energy-efficient systems, architected to deliver real-time inference will be critical for widespread deployment. The new NVIDIA GB200 NVL72 is one such system up to the task.

To illustrate, let's consider the Mixture of Experts (MoE) models. These models help distribute the computational load across multiple experts and train across thousands of GPUs using model parallelism and pipeline parallelism. Making the system more efficient.

However, a new level of parallel compute, high-speed memory, and high-performance communications could enable GPU clusters to make the technical challenge tractable. The NVIDIA GB200 NVL72 rack-scale architecture achieves this goal, which we detail in the following post.

A rack scale design for an exascale AI supercomputer

The heart of the GB200 NVL72 is the NVIDIA GB200 Grace Blackwell Superchip. It connects two high-performance NVIDIA Blackwell Tensor Core GPUs and the NVIDIA Grace CPU with the NVLink-Chip-to-Chip (C2C) interface that delivers 900 GB/s of bidirectional bandwidth. With NVLink-C2C, applications have coherent access to a unified memory space. This simplifies programming and supports the larger memory needs of trillion-parameter LLMs, transformer models for multimodal tasks, models for large-scale simulations, and generative models for 3D data.

The GB200 compute tray is based on the new NVIDIA MGX design. It contains two Grace CPUs and four Blackwell GPUs. The GB200 has cold plates and connections for liquid cooling, PCIe gen 6 support for high-speed networking, and NVLink connectors for the NVLink cable cartridge. The GB200 compute tray delivers 80 petaflops of AI performance and 1.7 TB of fast memory.

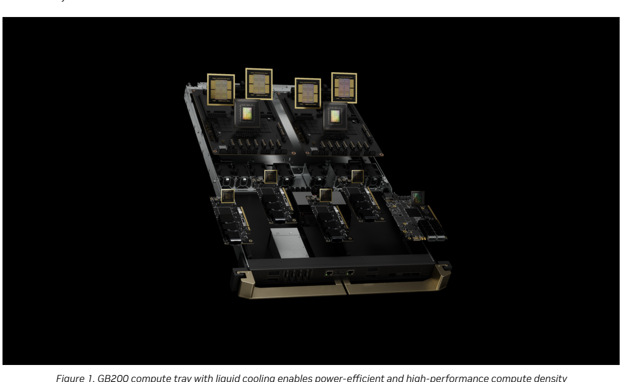


Figure 1. GB200 compute tray with liquid cooling enables power-efficient and high-performance compute density

The largest problems require sufficient numbers of groundbreaking Blackwell GPUs to work in parallel efficiently, so they must communicate with high bandwidth and low latency and keep constantly busy.

The GB200 NVL72 rack-scale system facilitates parallel model efficiency for the 18 compute nodes using the NVIDIA NVLink Switch System with nine NVLink switch trays, and cable cartridges interconnecting the GPUs and switches.

NVIDIA GB200 NVL36 and NVL72

The GB200 supports 36 and 72 GPUs in NVLink domains. Each rack hosts 18 compute nodes based upon the MGX reference design and the NVLink Switch System. It comes in the GB200 NVL36 configuration with 36 GPUs in one rack and 18 single GB200 compute nodes. The GB200 NVL72 is configured with 72 GPUs in one rack and 18 dual GB200 compute nodes—or 72 GPUs in two racks with 18 single GB200 compute nodes.

The GB200 NVL72 densely packs and interconnects the GPUs using a copper cable cartridge for operational simplicity. It also uses a liquid cooling system design, delivering 25x lower cost and energy consumption.

Figure 2. NVLink Switch System

Fifth-generation NVLink and NVLink Switch System

The NVIDIA GB200 NVL72 introduces fifth-generation NVLink, which connects up to 576 GPUs in a single NVLink domain with over 1 PB/s total bandwidth and 240 TB of fast memory. Each NVLink switch tray delivers 144 NVLink ports at 100 GB so the nine switches fully connect each of the 18 NVLink ports on every one of the 72 Blackwell GPUs.

The revolutionary 1.8 TB/s of bidirectional throughput per GPU is over 14x the bandwidth of PCIe Gen5, providing seamless high-speed communication for today's most complex large models.

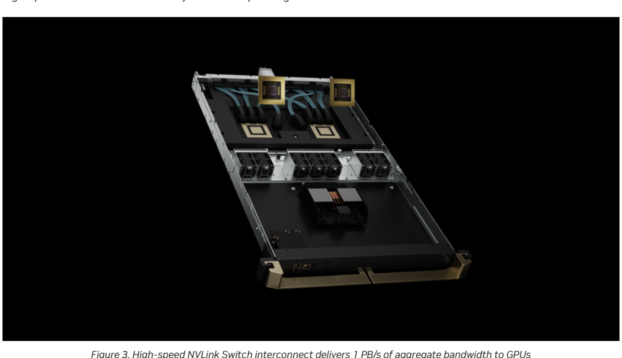


Figure 3. High-speed NVLink Switch interconnect delivers 1 PB/s of aggregate bandwidth to GPUs

NVLink over generations

The NVIDIA industry-leading innovation for high-speed low-power SerDes drives the advance of GPU-to-GPU communication, beginning with the introduction of NVLink to accelerate multi-GPU communications at high speed. The NVLink GPU-to-GPU bandwidth is 1.8 TB/s, which is 14x the bandwidth of PCIe. The fifth-generation NVLink is 12x faster than the first generation at 160 GB/s, introduced in 2014. NVLink GPU-to-GPU communications have been instrumental in scaling multi-GPU performance in AI and HPC.

GPU bandwidth advances coupled with the exponential expansion of NVLink domain size have increased the total bandwidth of an NVLink domain 900x since 2014 to 1 PB/s for a 576 Blackwell GPU NVLink domain.

Use cases and performance results

The compute and communications capabilities of the GB200 NVL72 are unprecedented, bringing grand challenges in AI and HPC within practical reach.

AI training

The GB200 includes a faster second-generation transformer engine featuring FP8 precision. It delivers 4X faster training performance with 32k GB200 NVL72 for large language models like GPT-MoE-1.8T compared to the same number of NVIDIA H100 GPUs.

AI inference

The GB200 introduces cutting-edge capabilities and a second-generation transformer engine that accelerates LLM inference workloads. It delivers a 30x speedup for resource-intensive applications like the 1.8T parameter GPT-MoE compared to the previous H100 generation. This advancement is made possible with a new generation of Tensor Cores, which introduce FP4 precision and the many advantages that come with fifth-generation NVLink

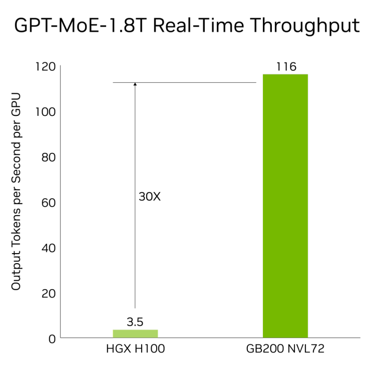


Figure 4. The GB200 delivers 30x real-time throughput compared to the H100

Results based on token-to-token latency = 50 ms; real-time, first token latency = 5,000 ms; input sequence length = 32,768; output sequence length = 1,024 output, 8x eight-way HGX H100 air-cooled: 400 GB IB Network vs 18 GB200 Superchip liquid-cooled: NVL36, per GPU performance comparison. Projected performance is subject to change.

The 30x speedup compares 64 NVIDIA Hopper GPUs scaled over 8-way NVLink and InfiniBand compared to 32 Blackwell GPUs in GB200 NVL72 using GPT-MoE-1.8T.

Data processing

Big data analytics helps organizations unlock insights and make better-informed decisions. Organizations continuously generate data at scale and rely on various compression techniques to alleviate bottlenecks and save on storage costs. To process these datasets efficiently on GPUs, the Blackwell architecture introduces a hardware decompression engine that can natively decompress compressed data at scale and speed up analytics pipelines end-to-end. The decompression engine natively supports decompressing data compressed using LZ4, Deflate, and Snappy compression formats.

The decompression engine speeds up memory-bound kernel operations. It provides performance up to 800 GB/s and enables Grace Blackwell to perform 18x faster than CPUs (Sapphire Rapids) and 6x faster than NVIDIA H100 Tensor Core GPUs for query benchmarks.

With the blazing 8 TB/s of high memory bandwidth and the Grace CPU high-speed NVlink-Chip-to-Chip (C2C), the engine speeds up the entire process of database queries. This results in top-notch performance across data analytics and data science use cases. This enables organizations to obtain insights quickly while reducing their costs.

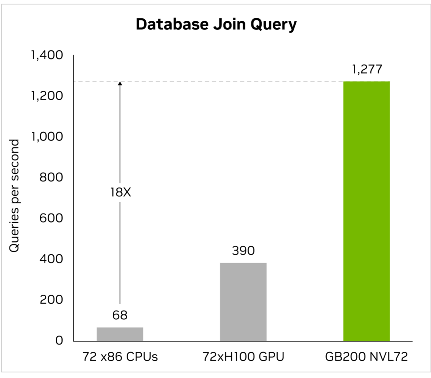


Figure 5. Database Join query throughput comparing GB200 NVL72, 72x H100, and 72 x86 CPUs

Physics-based simulations

Physics-based simulations are still the mainstay of product design and development. From planes and trains to bridges, silicon chips, and even pharmaceuticals, testing and improving products through simulation saves billions of dollars.

Application-specific integrated circuits are designed almost exclusively on CPUs in a long and complex workflow, including analog analysis to identify voltages and currents. The Cadence SpectreX simulator is one example of a solver. The following figure shows SpectreX runs 13x quicker on a GB200 than on a x86 CPU.

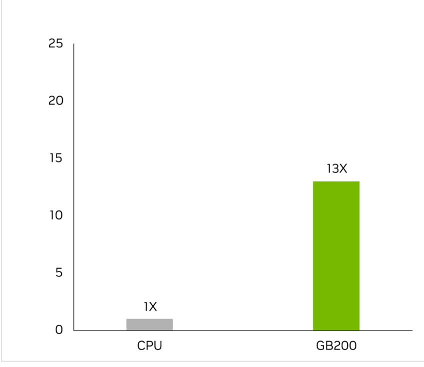


Figure 6. Cadence SpectreX simulator runs 13x quicker on a GB200 than on an x86 CPU

Cadence SpectreX (Spice simulator) | CPU: 16 cores AMD Milan 75F3 dataset: KeithC Design TSMC N5 | Performance projections for GB200 subject to change

Over the last two years, the industry has increasingly turned to GPU-accelerated computational fluid dynamics (CFD) as a key tool. Engineers and equipment designers use it to study and predict the behavior of their designs. Cadence Fidelity, a large eddy simulator (LES) runs simulations up to 22x faster on GB200 than x86 CPU.

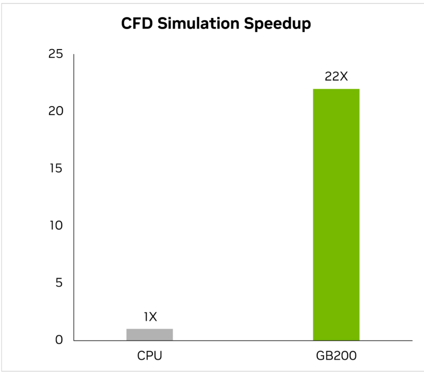


Figure 7. Cadence Fidelity, a large eddy simulator for CFD applications, runs simulations up to 22x faster on GB200 than x86 CPU

Cadence Fidelity (LES CFD Solver)| CPU: 16 cores AMD Milan 75F3 Dataset: GearPump 2M cells | Performance projections for GB200 subject to change

We look forward to exploring the possibilities of Cadence Fidelity on the GB200 NVL72. With parallel scalability and 30 TB of memory per rack, we aim to capture flow details that have never been captured before.

Summary

To recap, we reviewed the GB200 NVL72 rack-scale design and, in particular, learned about its unique capability to connect 72 Blackwell GPUs over a single NVIDIA NVLink domain. This reduces the communications overhead experienced when scaling over traditional networks. As a result, real-time inference for a 1.8T parameter MoE LLM is possible and training that model is 4x faster.

The operation of 72 NVLink-connected Blackwell GPUs with 30 TB of unified memory over a 130 TB/s compute fabric creates an exaFLOP AI supercomputer in a single rack. That is NVIDIA GB200 NVL72.

Related resources

- **GTC session:** LLM Inference Sizing: Benchmarking End-to-End Inference Systems
- **GTC session:** Enable Hybrid Training and Inference With DGX Cloud and OCI GPU Infrastructure (Presented by Oracle)
- **GTC session:** Speeding up LLM Inference With TensorRT-LLM: A Q&A From the EMEA Region
- **NGC Containers:** NVIDIA K8s Developer LLM Operator
- **NGC Containers:** NVIDIA MLPerf Inference
- **NGC Containers:** NVIDIA MLPerf Inference

Discuss (5) +28 Like

Tags

Data Center / Cloud | Generative AI / LLMs | Simulation / Modeling / Design | HPC / Scientific Computing | Grace CPU | NVLink | NVSwitch | Intermediate Technical | Deep Dive | AI Inference | Data Analytics / Processing | Featured | News | Physics | Technical Walkthrough

About the Authors

About Ivan Goldwasser

Ivan Goldwasser is a senior product marketing manager for the NVIDIA Data Center group, focusing on deep learning, training, and inference. Ivan has an MBA from Georgetown's McDonough School of Business and a bachelor's degree in chemical engineering from Texas A&M University.

View all posts by Ivan Goldwasser

About Harry Petty

Harry Petty is a senior technical marketing manager for HPC and AI edge applications at NVIDIA. Previously, he was a principal engineer and marketing director at Cisco Systems where he brought SDN innovations to market for hybrid cloud, multitenant security, and data center application performance. Harry has an MBA from Booth Graduate School of Business and a BS in mathematics and computer science from the University of Dayton.

View all posts by Harry Petty

About Pradyumna Desale

As a product lead in Enterprise Product Group at NVIDIA, Pradyumna Desale spearheads the scale up NVLink systems that leverage NVIDIA cutting-edge technology to address biggest customer challenges. He started his career at NVIDIA as a design engineer and later led a global engineering team that optimized the performance and power of high-speed IOs in NVIDIA GPUs and SoCs. He holds an MS degree in Electrical Engineering from Pennsylvania State University, State College and an MBA from University of California, Berkeley.

View all posts by Pradyumna Desale

About Kirthi Devleker

Kirthi Devleker is a senior product marketing manager at NVIDIA for the data center group, focusing on launching next-gen enterprise GPUs for generative AI training and inference applications. Before NVIDIA, Kirthi led strategic activities for launching next-gen AI-powered IaaS offerings at DigitalOcean. Kirthi also spent over 10 years at MathWorks as a senior product manager focused on evolving artificial intelligence and signal processing products and as a global medical devices industry manager. Kirthi has a background in signal and image processing and holds an MSEE degree from San Jose State University.

View all posts by Kirthi Devleker

Comments

Notable Replies

March 20, 2024
pldchang

Can you clarify if the BW is 900GB/s or 1.8TB/s (see the quote below)? Is liquid immersion required for 200Gbps serdes, 1.8TB/s NVLink Gen5 to function? Thanks.
"The heart of the GB200 NVL72 is the NVIDIA GB200 Grace Blackwell Superchip. It connects two high-performance NVIDIA Blackwell Tensor Core GPUs and the NVIDIA Grace CPU with the NVLink-Chip-to-Chip (C2C) interface that delivers 900 GB/s of bidirectional bandwidth."

March 20, 2024
kiranpalli

I believe GPU-GPU NVLink BW is 1.8TB/s but as the Grace CPU is being reused from GH100, CPU-GPU NVLink BW is still 900GB/s.

March 20, 2024
kiranpalli

B100 and B200 seem to have 180GB GPU memory as per this article: NVIDIA HGX AI Supercomputing Platform
but GB200 with two GPUs seems to have 394GB (192GB per GPU) of GPU memory. Wondering if there is a typo somewhere.

March 21, 2024
pldchang

Hi Kiranpalli,
Thanks for the clarification. Are the NVLink 5th gen 1.8TB Cu cables the same as those of NVLink gen 4 (0.9TB)? Thanks.
Best,
Peter

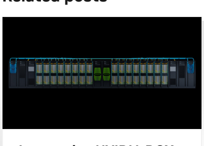
March 21, 2024
hpetty

pldchange, NVLink-C2C is a 900GB/s bidirectional interconnect between CPU and GPU, used for Grace and Hopper or Blackwell GPU(s). In the case of GB200, it is shared between two Blackwell GPUs and Grace. 5th generationNVLink is a1.8TB/s interconnect between GPUs. Liquid cooling is used to dissipate the heat energy from the Superchip but not related to the SerDes to function. The 5th generation NVLink cables are spec'd for the fast 1.8TB/s data rate.

Continue the discussion at forums.developer.nvidia.com

Participants

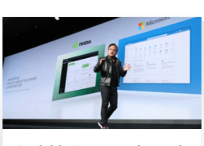
Related posts



Announcing NVIDIA DGX GH200: The First 100 Terabyte GPU Memory...

Scaling Out the Deep Learning Cloud Efficiently

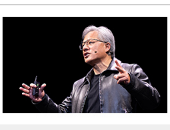
OpenAI Presents GPT-3, a 175 Billion Parameters Language Model



Scalable GPU-Accelerated Supercomputer in the Microsoft Azure Cloud

MIT Lincoln Laboratory Supercomputing Center Installs World's Fastest...

Don't Miss This Transformative Moment in AI
Live On-Stage Keynote
Watch Now



Jensen Huang
Founder & CEO, NVIDIA
NVIDIA GTC

Copyright © 2024 NVIDIA Corporation